Fuzzy-TLX: using fuzzy integrals for evaluating human mental workload with NASA-Task Load indeX in laboratory and field studies

Marc Mouzé-Amady[a]*, Eric Raufaste[b], Henri Prade[c] and Jean-Pierre Meyer[a]

[a]*Occupational Physiology Laboratory , Institut National de Recherche et de Sécurité, 1 rue du Morvan, CS 60027, F-54519 Vandœuvre, France;* [b]*CLLE-LTC, CNRS, Université de Toulouse, EPHE, 5, Allées A. Machado, F-31058 Toulouse Cedex 9, France;* [c]*IRIT, CNRS, Université de Toulouse, 118, route de Narbonne, F-31062 Toulouse Cedex 9, France*

The aim of this study was to assess mental workload in which various load sources must be integrated to derive reliable workload estimates. We report a new algorithm for computing weights from qualitative fuzzy integrals and apply it to the National Aeronautics and Space Administration -Task Load indeX (NASA-TLX) subscales in order to replace the standard pair-wise weighting technique (PWT). In this paper, two empirical studies were reported: (1) In a laboratory experiment, age- and task-related variables were investigated in 53 male volunteers and (2) In a field study, task- and job-related variables were studied on aircrews during 48 commercial flights. The results found in this study were as follows: (i) in the experimental setting, fuzzy estimates were highly correlated with classical (using PWT) estimates; (ii) in real work conditions, replacing PWT by automated fuzzy treatments simplified the NASA-TLX completion; (iii) the algorithm for computing fuzzy estimates provides a new classification procedure sensitive to various variables of work environments and (iv) subjective and objective measures can be used for the fuzzy aggregation of NASA-TLX subscales.

Practitioner Summary: NASA-TLX, a classical tool for mental workload assessment, is based on a weighted sum of ratings from six subscales. A new algorithm, which impacts on input data collection and computes weights and indexes from qualitative fuzzy integrals, is evaluated through laboratory and field studies. Pros and cons are discussed.

Keywords: fuzzy integral; NASA-TLX; ageing; stress

1. Introduction

1.1. The NASA Task Load indeX

The National Aeronautics and Space Administration-Task Load indeX (NASA-TLX) has been extensively used in human performance studies to assess subjective workload: it has been translated into more than a dozen languages, and 82,900 citations were found when googling for 'NASA TLX' (Hart 2006).

In their 1988 seminal paper, Hart and Staveland (1988, 140) defined workload as 'a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. Thus, our definition of workload is human-centered, rather than task-centered'. Eighteen years later, a close definition was still maintained (Hart 2006). Since workload is multifaceted, NASA-TLX was designed as a six-dimension scale for estimating the workload felt by operators. The six visual analogue subscales used for rating are as follows: mental demand (MD), physical demand (PD) and temporal demand (TD), Frustration level (FL), Effort level (EL) and Performance level (PL). The overall workload estimate is computed as a weighted average of the six original ratings. In order to account for individual differences, each rating is weighted by its own coefficient issued from the pairwise weighting technique (PWT): the operator is asked, over 15 pairwise comparisons, to choose which dimension is the most important as a workload source (MD vs. PL, EL vs. FL and so on). The relative weight of each source is then computed by counting the number of comparisons where it ranked first, divided by 15 for normalisation purposes. Finally, a weighted sum provides the aggregated index (TLX):

$$\mathrm{TLX} = \mathrm{MD} * \mathrm{W_{MD}} + \mathrm{PD} * \mathrm{W_{PD}} + \mathrm{TD} * \mathrm{W_{TD}} + \mathrm{FL} * \mathrm{W_{FL}} + \mathrm{EL} * \mathrm{W_{EL}} + \mathrm{PL} * \mathrm{W_{PL}}$$

where W_{XX} is the weight of the rating made on the subscale *XX*. TLX can be rewritten in a short form as follows:

$$\mathrm{TLX} = \sum_{i=1}^{6} w_i a_i, \tag{1}$$

where w_i and a_i denote the weight and rating associated with the *i*th workload source.

1.2. Methodological and practical issues

PWT is easily calculable, but it raises some methodological and practical problems, especially in real work environments.

First, the weighted average is based on mathematical assumptions usually not verified. Thus, weights should be independent from ratings as when operators provide ratings and external experts independently provide weights. But it is not the case in the standard NASA-TLX procedure where each operator provides both ratings and weights. Another major limitation of PWT is the fact that it cannot take into account workload source interactions, although the effect of two (or more) load sources may not always combine additively. For example, there may be potentiating effects where the results of Loads A and B combine multiplicatively. Furthermore, classical associative rules may not apply: operators exposed to '(loads A and B) and (load C)' may react (rate) differently if exposed to '(load A) and (loads B and C)' and so on. Moreover, sometimes, the transitivity rules may be violated (e.g. if MD $>$ TD and TD $>$ EL, participants may not respond that MD $>$ EL) (Nygren 1991).

Second, pairwise comparisons may also be hard to be provided by low-qualified workers due to the abstraction level needed (e.g. 'Which was more important in your task: MD or FL?') as observed in some French field studies (Liévin and et François 1997). Third, the relevance of PWT versus raw TLX has also been called into question by others (Byers, Bittner, and Hill 1989). Fourth, the scale reliability has been discussed when some items are not associated with task demands: e.g. PD in office workers (Bridger and Brasher 2011).

Last but not the least, rating the six subscales and then performing PWT are time consuming, especially in work situations (not to mention the time needed to instruct operators on how to fill the questionnaire) and/or when questionnaire administration needs to be repeated.

1.3. A potential solution: the fuzzy integral approach to loads aggregation (Fuzzy-TLX)

Aggregation refers to the process of combining numerical values into a single new one. Generally, loads aggregation combines each rating with its weight and then uses some rule to aggregate intermediate results into the final value (Equation (1)). As a weighted average, PWT combines multiplicatively the ratings and their associated weights, then additively the intermediate results. Since completion of NASA-TLX may be viewed as how a respondent defines his/her own perception of a set representing a linguistic term, '*item*', describing a variable '*load*', by offering the respondent many different values of *load* (ratings), this data collection for measuring fuzzy membership sets was termed direct (D) continuous (C) method (Brackstone 2000). Previously, some authors proposed to use this DC method to feed the weight averaging process (Liu and Wang 1994; Chen 1996). This paper presents Fuzzy-TLX, a radically new method since it directly produces weights based on general multiple criteria aggregation devices, namely Sugeno integrals (Sugeno 1974, 1977).[1] These integrals are qualitative functions that can be defined on any completely ordered scale. They return the median of a set of values made from partial evaluations to be combined and of weights associated with subsets of these partial evaluations (Raufaste and Prade 2006; Prade et al. 2009). See Appendixes 1 and 2 for further mathematical and practical details.

The weights obtained in this way are not provided by participants, thus attenuating various measurement bias and practical problems. Fuzzy-TLX is qualitative in essence but can be extended to other multiple-criteria approaches using weighted sum aggregations (references *in* Raufaste and Prade 2006). Contrary to standard NASA-TLX using PWT, Fuzzy-TLX can handle interactions. Furthermore, it enables the detection of 'aggregation policies'. Each aggregation policy corresponds to a coherent set of weights. Environment changes can substantially modify the relative weights of load sources, generating a variety of aggregation policies.

To test these properties of Fuzzy-TLX, the data reported here were obtained from a laboratory study and a field study both addressing mental workload and stress. Only NASA-TLX data from both studies are presented here, from which two Fuzzy-TLX were computed, i.e. the maximum (SugMax) and minimum (SugMin) boundaries of the Sugeno integral. They correspond to high and low mental workload estimates.

2. The laboratory study

First-world countries face the demographic challenge of their ageing workforce. As part of a general study on ageing, health and work, a laboratory experiment (Gilles et al. 2007a, 2007b) was set up to investigate the functional capacities (cognitive and physical) of active or retired workers of various ages. We only report participants' NASA-TLX results issued from classical and fuzzy treatments.

2.1. Participants

Fifty-three healthy male volunteers were included after recruitment by a regional temporary employment agency. They were either active or retired manual workers from the industrial or building sectors. To avoid cultural influences on

performing NASA-TLX (Johnson and Widyanti 2011),[2] participants were all from the same national culture. Moreover, since computers are now commonly used at workplaces and at home, all participants were regular computer users. Upon completion of the experiment (three mornings), they received monetary compensation. For data analyses, they were split into three age brackets: 'young', 30–35 years, $n = 16$, 'medium', 45–50 years, $n = 19$ and 'senior', 60–65 years, $n = 18$.

2.2. Experimental tasks

The study was approved by the local Ethics Committee of the Nancy region (France). For each participant, the experimental procedure was conducted during three consecutive mornings in the INRS laboratories of occupational physiology. Morning 1 was devoted to medical examination, detailed information about the whole procedure, written consent from the participant and various functional tests (step test, oxygen consumption, grip strength and so on). On mornings 2 and 3, the participant performed two tasks: a computer task and a physical task. The computer task was a home-made computerised version of the classical Stroop colour word test[3] (Stroop 1935). In this experiment, a coloured word naming a colour was displayed on screen in random order and at random time intervals (machine-determined pace). Participants were instructed to answer as quickly and correctly as possible by pressing one of the two keyboard keys allocated to 'congruent' (e.g. 'RED' displayed in red) or 'not congruent' (e.g. 'BLUE' in green). The 150 stimuli display lasted about 15 min. The Stroop test is known to elicit conflict perception leading to a stress state while as a response to mental demand heart rate variability (HRV) is reduced over time (Hoshikawa and Yamamoto 1997). At the end of the Stroop task, participants were asked to rate their associated subjective mental workload. In the physical task, they had to mount an object in five steps: (1) take a wood base from an imposed speed rotating cylinder, (2) walk to the workbench, (3) grab two bolts and one handle placed under the workbench, (4) put all elements together, (5) bring back the mounted object to the display cylinder before it starts a new turn. This mounting task was repeated in loop during 20 min. Two mounting speed rates were imposed via the automated rotating cylinder: (1) 25 s by cylinder turn ('comfortable' condition [Cc]) and (ii) 20 s by cylinder turn ('rapid' condition [Rc]). Half of the participants performed Cc on Day 2 and Rc on Day 3. The order was reversed for the other 50%. At the end of the physical tasks (Cc and Rc), participants were asked to rate their associated subjective mental workload.

2.3. Workload measurement

Subjective measures of workload. After each task (Stroop test, Physical tasks), participants were asked to rate the subjective mental workload they had felt using a computerised version of NASA-TLX. The interface first provided the six NASA-TLX subscales plus a global subjective workload (GSW) scale added for further fuzzy processing. Each rating was done by moving a mouse-driven cursor on a graphical bounded line (0–10) and validated by clicking the 'Next' button. Then data for the PWT were collected through a series of 15 screens: each displaying a pair of items (MD vs. PD, TD vs. FL and so on). Participants had to click the item they felt the most salient during the task.

Objective measures of workload. During the tasks, participants' cardiac signals (interbeat intervals) were continuously recorded by portable cardiofrequency meter (Polar®) for further HRV analyses with Kubios HRV 2.0 software (University of Kuopio 2008). A nonlinear HRV parameter called 'recurrence rate of consecutive heartbeats' (%REC) was chosen (for Recurrence Quantification Analysis (RQA),[4] see Webber and Zbilut 1994; Zbilut, Thomasson, and Webber 2002). Being a percentage, it enables easy adjustment to the boundaries of the NASA-TLX subscales and thus could serve as a criterion for fuzzy aggregation. Moreover, HRV is related to changes in sympathovagal balance related to psychological or environmental stressors. It is noteworthy for result interpretation that stress increases %REC values (review *in* Von Borell et al. 2007).

Subjective measures and derived analyses. Classical TLX values were computed using individual PWT data, whereas Sugeno integrals were computed over all NASA-TLX subscales. Two types of fuzzy results were obtained depending on the subjective or objective nature of the aggregating criterion (GSW or %REC). For each Sugeno integral, both the lower (SugMin) and upper (SugMax) boundaries were introduced in non-parametric ANOVAs. Interactions on ranks were estimated using the Scheirer–Ray–Hare test (Sokal and Rohlf 1995).

2.4. Results

2.4.1. Psychometric analyses

Fuzzy measures derived from GSW. From the three groups (young, medium and senior), 156 valid[5] NASA-TLX out of 159 were selected. Overall spearman correlations between TLX and Sugeno integrals were $\rho(156) = 0.833$, $p < 0.001$ and $\rho(156) = 0.828$, $p < 0.001$ for the lower (SugMin) and upper (SugMax) boundaries, respectively. Detailed ρ_s for the three age brackets times the three tasks ranged from 0.58 to 0.96, all being statistically significant ($p < 0.05$), especially in young

and medium groups ($p < 0.001$). To test Fuzzy-TLX reliability or internal consistency, Cronbach's α was computed for the ratings in each subgroup and for both fuzzy scores (SugMax and SugMin). All coefficients exhibited good reliability of the modified NASA-TLX ($\alpha > 0.75$).

Fuzzy measures derived from HRV (%REC). In the following results, weights were computed from HRV as indexed by the recurrence rate (%REC). Using %REC resulted in a 'loss' of data ($n = 119$ with %REC), thereby reducing statistical power. Spearman rank correlations between NASA-TLX and Fuzzy-TLX were $\rho(119) = 0.374$, $p < 0.001$ for SugMin and $\rho(119) = 0.391$, $p < 0.001$ for SugMax.

Summary of psychometric analyses. The psychometric properties described above reflect good internal consistency between the six workload dimensions and Sugeno integrals. Besides, clear but not perfect correlations was found between NASA-TLX and Fuzzy-TLX, showing that standard NASA-TLX and Fuzzy-TLX are different even though sharing some similarity. The fact that similarity decreases in some groups might be interpreted in two contradictory ways. Either groups differ on workload and NASA-TLX lack discrimination or groups do not actually differ and Fuzzy-TLX misreports differences (false alarms). But following results about task and ageing suggest that such differences actually exist.

2.4.2. *Effects of task and ageing*

Fuzzy measures based on GSW. Figure 1 exhibits a slight descriptive difference between NASA-TLX and Fuzzy-TLX capacities to discriminate between the physical Rc and Stroop tasks. For NASA-TLX, non-parametric ANOVA discriminates the tasks ($H_{(2,155)} = 16.11$, $p < 0.001$) but neither age nor the interaction (age × task). Post hoc tests show that NASA-TLX values for Rc and Stroop were not significantly different ($z = 0.212$, *ns*), whereas for both SugMax and SugMin, the difference approaches significance ($z = 1.66$, $p = 0.097$ and $z = 1.776$, $p = 0.076$).

Fuzzy measures based on %REC. First, mean %REC was higher in medium (42.74 ± 1.37) and senior (42.45 ± 1.41) groups, than in young (37.93 ± 1.54), non-parametric ANOVA: $H_{(2,117)} = 6.52$, $p = 0.038$. This is in accordance with the literature (Gregoire et al. 1996; Antelmi et al. 2004)[6] describing a decrease in HRV with age. Second, Figure 2 shows how the three workload indexes were affected by the task. For NASA-TLX, only the task factor was significant: $H_{(2,118)} = 10.59$, $p = 0.005$. For the fuzzy estimates (SugMax and SugMin), the task factor was also significant: $H_{(2,118)} = 13.23$, $p = 0.001$ and $H_{(2,118)} = 16.37$, $p < 0.001$, respectively. Age and interaction (age × task) were non-significant for SugMax, but for SugMin age approached significance: $H_{(2,118)} = 4.73$, $p = 0.094$ (Figure 3).

The use of an objective criterion (%REC) as a seventh item for fuzzy aggregation in place of the subjective criterion (GSW) led to a sharper 'cut-off' threshold in original data (119 vs. 156). In order to compare both criteria, data 'hold back' by the %REC criterion were reprocessed with the subjective criterion. Table 1 summarises rank correlations between NASA-TLX and FUZZY-TLX, and Cronbach's α for NASA-TLX subscales + criteria (GSW and %REC).

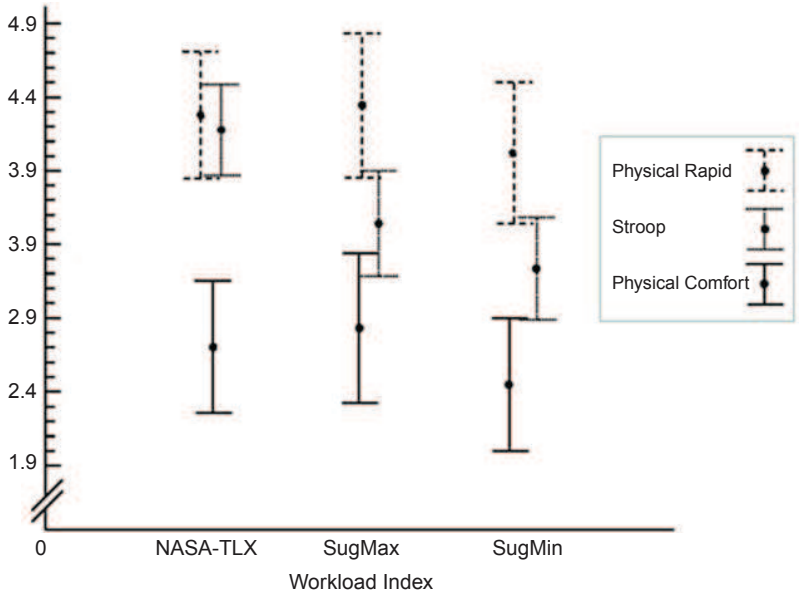

Figure 1. Mean workload values with 95% confidence intervals for NASA-TLX and the two fuzzy values computed from subjective-data-derived weights, in three conditions: physical Comfort condition, physical Rapid condition and Stroop test.

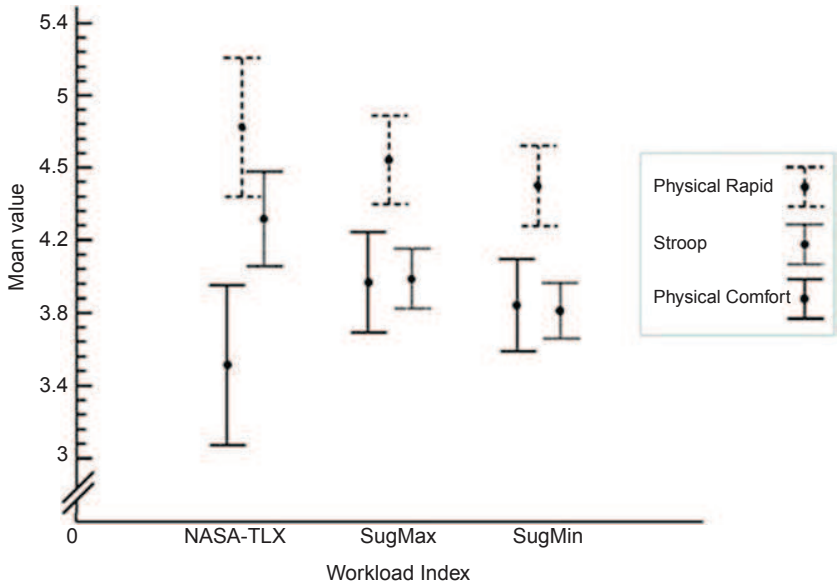

Figure 2. Mean workload values with 95% confidence intervals for NASA-TLX and the two fuzzy values computed from objective-data-derived weights, in three conditions: physical Comfort condition, physical Rapid condition and Stroop test.

A majority of Cronbach coefficients (25 out of 36) suggest good internal consistency between the six ratings and the aggregated fuzzy values ($\alpha > 0.70$), whereas a majority of non-significant correlations with NASA-TLX appear (27 out of 36).

Regarding tasks and age factors, GSW-based fuzzy estimates provided no significant effect, in contrast to %REC-based fuzzy estimates (see above and Figure 3). Thus, the objective criterion (%REC) seems more powerful to single out variable effects on a reduced data sample than the subjective criterion (GSW). Moreover, looking for consistent subsets of Sugeno integrals based on the %REC criterion, a partition of five data subsets was returned by the algorithm: #1 ($n = 73$), #2 ($n = 15$), #3 ($n = 25$), #4 ($n = 4$) and #5 ($n = 2$). Each subset represents a particular aggregation policy, but #4 and #5 were ignored due to data scarcity. A discriminant analysis was performed on the three main subsets (i.e. 113 values out of 119) with age as clustering variable. Two discriminant functions were found (Figure 4). Function 1 discriminates senior and young groups (negative part of the graph) from the Medium group (positive part), $\chi^2_{(6)} = 21.97$, $p < 0.001$; Function 2 discriminates young and medium groups (negative part) from the senior group (positive part), $\chi^2_{(2)} = 8.14$, $p < 0.05$. This result highlights the classification ability of Sugeno estimates.

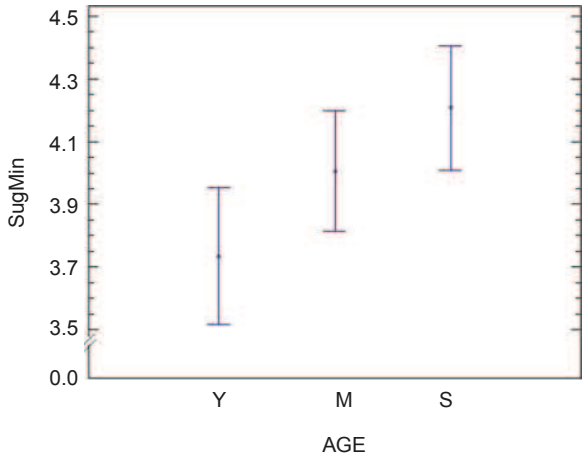

Figure 3. Mean SugMin with 95% confidence interval for the three age brackets: young (30–35 years), medium (45–50 years) and senior (60–65 years). Significant level: $p = 0.094$.

Table 1. Spearman rank correlations (ρ) between TLX and Sugeno integral boundaries (based on subjective (GSW) or objective (% REC) data-derived weights) and Cronbach's α, within Age × Task groups (significant values in bold).

		Young				Medium				Senior			
Task	Index	ρ	N	Sig(ρ)	α	ρ	N	Sig(ρ)	α	ρ	N	Sig(ρ)	α
Physical Cc	SugMinGSW	0.61	8	*ns*	**0.87**	−0.29	9	*ns*	0.59	−0.26	6	*ns*	**0.83**
	SugMaxGSW	0.52	8	*ns*	**0.87**	−0.030	9	*ns*	0.57	−0.26	6	*ns*	**0.82**
	SugMin%REC	0.55	8	*ns*	**0.82**	0.59	9	<0.10	0.63	0.67	6	*ns*	**0.79**
	SugMax%REC	**0.79**	8	<0.05	**0.83**	0.59	9	<0.10	0.62	0.71	6	*ns*	**0.79**
Physical Rc	SugMinGSW	0.55	9	*ns*	**0.85**	0.46	9	*ns*	0.85	0.43	12	*ns*	**0.83**
	SugMaxGSW	0.53	9	*Ns*	**0.84**	0.46	9	*ns*	**0.84**	0.35	12	*ns*	**0.82**
	SugMin%REC	0.29	9	*ns*	**0.74**	0.10	9	*ns*	**0.79**	−0.13	12	*ns*	**0.72**
	SugMax%REC	0.39	8	*ns*	**0.74**	0.10	9	*ns*	**0.79**	−0.09	12	*ns*	**0.74**
Stroop	SugMinGSW	0.29	17	*ns*	**0.86**	**0.66**	26	<0.001	**0.80**	**0.55**	23	< 0.01	0.69
	SugMaxGSW	0.28	17	*ns*	**0.86**	**0.61**	26	<0.001	**0.80**	**0.53**	23	< 0.01	0.69
	SugMin%REC	**0.67**	17	<0.01	**0.80**	**0.42**	26	<0.05	0.67	0.03	23	*ns*	0.53
	SugMax%REC	**0.55**	17	<0.05	**0.80**	**0.50**	26	<0.01	0.68	0.00	23	*ns*	0.53

3. Field study

Nowadays, a majority of air transport operations are commercial short and medium hauls. Although working conditions and health effects on long and very long hauls are well documented in the scientific literature, this is not the case for short or medium hauls (bibliometric analyses *in* François et al. 2007). Vis-a-vis this gap, INRS was asked by people in charge of safety and health of airline crews to carry out a study on the effects of working conditions in these types of flights.

3.1. Design and participants

This study targeted the human–environment system and its effects on workload and stress by means of interviews, behavioural observations, workload scales, physiological and environmental measures during 48 *real* commercial flights (see footnote) with volunteer crews. Stress and working conditions were also assessed by means of questionnaires (details *in* François et al. 2007). In this paper, we only report pilots' NASA-TLX results issued from fuzzy treatments during commercial flights (for cabin crew results, see Raufaste and Prade 2006). Ratings were collected during five aircraft rotations from a big European airline: each rotation lasted 3 days with three, four and three successive flights per day, respectively. During each rotation (10 flights),[7] the same flying personnel was maintained (which is not the case in most common rotations) for methodological purposes. Ten pilots (five Captains and five First Officers) responded to the NASA-TLX questionnaire (i.e. 380 raw NASA-TLX).

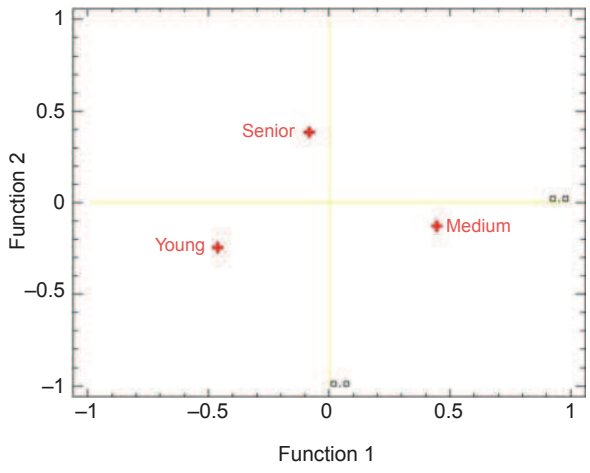

Figure 4. Graph of the discriminant functions from Sugeno integrals based on %REC criterion with the three age centroids: young (30–35 years), medium (45–50 years) and senior (60–65 years).

3.2. *Materials*

A modified NASA-TLX paper version included the six classical items (MD, PD, TD, FL, EL and PL subscales) and a seventh item for GSW ('How much was your global workload to perform the task?'), all ranging from low(0) to high(10).

3.3. *Procedure*

Each flight included four main time phases: preparation, take-off, cruise and landing. For practical and safety reasons, the subjective evaluation covered several activities within each main phase. Preparation phase included flight plans, briefing between the whole aircrew, out and indoor security checks of the airliner, (re)fuelling, catering (meal on wheels delivery), passengers boarding and passengers headcount. Take-off phase included to secure passengers and deliver safety advices, monitoring the pushback of the aircraft from apron, taxiing out on runway, final take-off and initial climb. Cruise phase included food and commercial services and navigation operations. Landing phase included to secure passengers, final approach, landing, taxiing in, deboarding and transit time.

For each phase, pilots had to rate their subjective load level on the seven subscales. Due to time constraints associated with real flight conditions, PWT was not performed.

3.4. *Results*

Concerning the flight deck crews (Captain and First Officer), fuzzy measures of mental workload are reported in relation with job qualification and flight phases. But first, the reliability or internal consistency of the modified NASA-TLX was tested by means of Cronbach's α (computed for the six classical ratings and the fuzzy scores within the four flight phases, Table 2).

Such coefficients show a relative good reliability (>0.70) of the Fuzzy-TLX. Table 3 shows, for both pilots, Spearman rank correlations between GSW ratings and Sugeno estimates based on the six classical NASA-TLX ratings.

Despite both pilots using the same instruments and sharing the same goal (to safely transport passengers), fuzzy estimates showed that their respective mental workload evaluations were quite different. Both high (SugMax) and low (SugMin) levels of subjective mental workload can be distinguished as job (Figure 5 and Table 4) and flight phase related (Figures 6(a) and (b)). There was no significant interaction between the two variables.

Overall, (a) subjective mental workload estimates were significantly higher in Captains than in co-pilots (Figure 5 and Table 4), (b) the time course of the estimates was similar in both pilots (Figure 6(a),(b), (c) critical phases (preparation, take-off and landing) were reflected by higher values, (d) the lowest values were observed during the cruise phase and (e) the subjective 'mental bandwith' (SugMax minus SugMin) was very narrow since SugMax and SugMin values were very close (Figure 5).

4. Discussion

An original algorithm able to extract a qualitative tool called Sugeno integral was applied to compute a fuzzy analogue of NASA-TLX, bypassing the classical PWT. At the mathematical level, Fuzzy-TLX is better grounded than NASA-TLX as it avoids a series of flaws such as violations of weight-rating independence. It also allows a genuinely qualitative treatment of subjective workload measures. Using only one external[8] (subjective or objective) criterion, it can determine the ratings' weights. At the psychometric level, the method is in accordance with the international standard requirements (ISO 10075-3 2005) for measuring subjective mental workload. Moreover, it enables collecting weights in a much less time-consuming fashion more compatible with real work environments. This paper reported testing of empirical properties of this new tool for subjective workload assessment.

Table 2. Cronbach's α for SugMax (up right corner) and SugMin (down left corner) within the four flight phases and for both pilots, n is equal to the number of valid-filled NASA-TLX forms per phases during the five rotations.

Job	Preparation	Take-off	Cruise	Landing
Captain ($n = 47$/phases)	0.74	0.73	0.90	0.78
	0.75	0.72	0.90	0.78
Copilot ($n = 48$/phases)	0.79	0.72	0.78	0.72
	0.79	0.72	0.78	0.72

Table 3. Spearman's rho (ρ) between GSW and Sugeno's integrals in Captains and First Officers, n is equal to the number of valid-filled NASA-TLX forms during the five rotations.

	GSW
SugMin	
Captains ($n = 188$)	0.94*
First Officers ($n = 192$)	0.93*
SugMax	
Captains ($n = 188$)	0.94*
First Officers ($n = 192$)	0.93*

Note: * denotes significance level at $p < 0.05$.

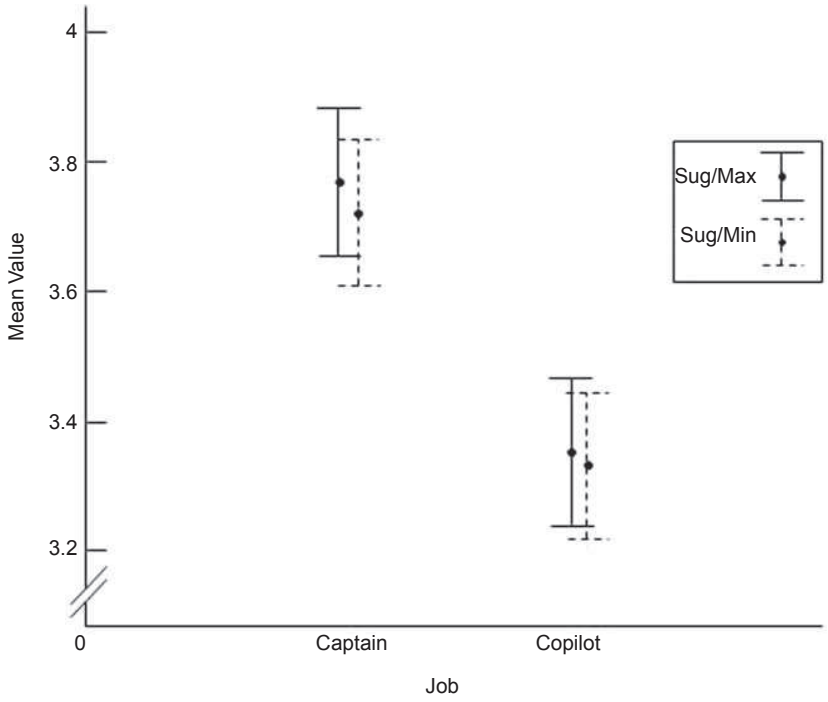

Figure 5. Mean values with 95% confidence interval of SugMax and SugMin values for Captains and First Officers' subjective mental workloads.

Table 4. Kruskal–Wallis ANOVA (and Scheirer–Ray–Hare test for interaction) with two factors: type of job (Captain or Copilot) and time phases (preparation, take-off, cruise or landing).

Source	df	H	p-value
SugMax ($n = 384$)			
Type of job (A)	1	18.09	0.00002**
Time phases (B)	3	29.43	1.821e − 06**
Interaction (A × B)	3	1.73	0.63098
SugMin ($n = 384$)			
Type of job (A)	1	15.94	0.00007**
Time phases (B)	3	27.34	5.007e − 06**
Interaction (A × B)	3	2.21	0.53034

Note: ** denotes significance level at $p < 0.01$.

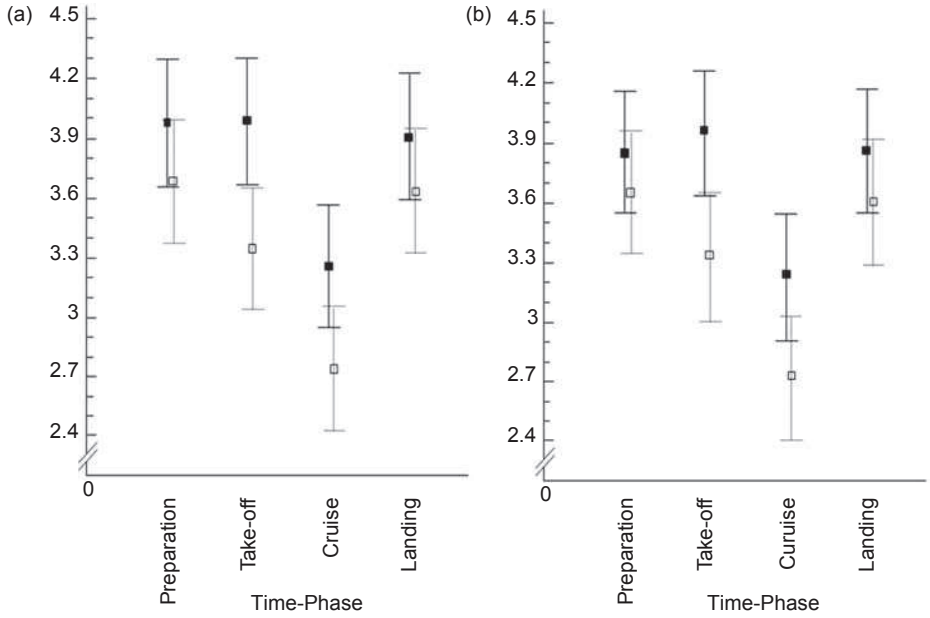

Figure 6. (a). Mean values with 95% confidence interval of the upper (SugMax) boundary for Captains and Copilots' subjective mental workloads during the four time phases (■, Captains; □, Copilots). (b) Mean values with 95% confidence interval of the lower (SugMin) boundary for Captains and Copilots' subjective mental workloads during the 4 time-phases (■, Captains; □, Copilots).

In the laboratory study, the main variables under study were age (young, medium and senior participants) and type of task (two physical tasks and a cognitive test). Besides computer-based NASA-TLX subscales, two criteria were collected: GSW and %REC (a nonlinear index of HRV). Each allowed computation of both minimum and maximum fuzzy integral boundaries: SugMin and SugMax. As said earlier (§2.1.), participants were all regular computer users, and no work overload[9] was associated with the media contrary to other studies (e.g. Noyes and Bruneau 2007).

For GSW-based Fuzzy-TLX, overall data as well as within subgroups (age and task), the correlations between classical TLX (with PWT) and fuzzy estimates were statistically significant. Fuzzy-TLX reliability was reflected by high value (>0.70) of the Cronbach's α issued from the six classical ratings and the fuzzy scores. Furthermore, the discrimination between the task variables was higher with Fuzzy-TLX than with NASA-TLX. However, both methods failed to detect an ageing effect. With HRV (indexed by %REC) as the aggregation criterion, at least one estimate (the lower bound of Fuzzy-TLX) was able to detect both ageing and task effects. This finding shows that fuzzy devices can 'mix' subjective and objective data in order to extract more information from the situation under study. This result is in accordance with another study (in air traffic control) in which the authors argued for the need to mix objective measures (conflict times in their case) with NASA-TLX data (Averty et al. 2004). In our case (fuzzification of mixed data) would the Choquet integral (quantitative counterpart of Sugeno integral) give better results? The question remains open. However, it should be remarked that provided that some software development be done, nothing prevents using Choquet integral in the same way as we did with qualitative data.

In the field study (on flying personnel during real flights), the PWT was not performed due to lack of time and binding security rules, so Fuzzy-TLX was applied alone. As for cabin crews (Raufaste and Prade 2006), the fuzzy treatments based on GSW were able to discriminate job- and task-related variables with significant psychometric properties in flight deck crews. It was observed that the subjective mental workload estimates were higher in Captains than in Copilots, probably reflecting differences in flying responsibilities (even if the time course of these estimates was similar in both pilots). Moreover, higher values of mental workload were found during the most critical flight phases: preparation, take-off and landing. This result was in accordance with verbal reports (François et al. 2007). On the other hand, the lowest values of workload were observed during the cruise phase where flying control is left to the auto pilot. We also observed that the subjective 'mental bandwidth' as defined by the boundaries of Sugeno integrals (SugMax and SugMin) was very narrow; this may be related to the very high level of technical flying procedures (reduced degrees of freedom). Unfortunately, the comparison between classical NASA-TLX and fuzzy estimates was not available as previously said.

5. Conclusion

Overall, in both studies (laboratory and field), results demonstrate that Fuzzy-TLX makes it possible to collect valuable information on subjective workload in various settings and to assess mental workload where original NASA-TLX cannot be used.

Acknowledgement

The authors acknowledge Jean-Charles Guélin[a] for technical assistance in both studies.

Notes

1. There are at least two fuzzy integrals: the Sugeno integral and the Choquet integral (Murofushi and Sugeno 1991). Briefly, the first one best applied on qualitative data, whereas the second one on quantitative data.
2. Note that authors used only unweighted NASA-TLX ratings.
3. On day 1, a screening for colour blindness was done using the Ishihara chromatic test.
4. Contrary to time- or frequency-domain analyses, RQA is independent of time series size, stationarity or specific statistical distribution of data.
5. Some participants failed to rate coherent scores or failed during the PWT.
6. These studies, however, report time- and frequency-domain indexes rather than chaotic measures such as %REC.
7. 10 flights × 5 rotations = 50 flights, but 2 flights were cancelled during the whole study.
8. Meaning: not part of original NASA-TLX subscales.
9. Except for the completion of the PWT itself (i.e. systematic verbal complaints).
10. The algorithm computes not only 6 but also 62 weights ($2^6 - 2$: since $\mu(\varnothing) = 0$ and $\mu(N) = 1$), thus providing interaction weights.
11. See (1) and (2) in Appendix 1.
12. {123456} is not relevant, here, since its weight is always 100 (see (2) in Appendix 1).
13. The outlier observation is then rejected.
14. See (3) in Appendix 1.

References

Antelmi, I., R. S. de Paula, A. R. Shinzato, C. A. Peres, A. J. Mansur, and C. J. Grupi. 2004. "Influence of Age, Gender, Body Mass Index, and Functional Capacity on Heart Rate Variability in a Cohort of Subjects without Heart Disease." *The American Journal of Cardiology* 93 (3): 381–385.

Averty, P., C. Collet, A. Dittmar, S. Athènes, and E. Vernet-Maury. 2004. "Mental Workload in Air Traffic Control: An Index Constructed from Field Tests." *Aviation Space and Environmental Medicine* 75 (4): 333–341.

Brackstone, M. 2000. "Examination of the Use of Fuzzy Sets to Describe Relative Speed Perception." *Ergonomics* 43 (4): 528–542.

Bridger, R. S., and K. Brasher. 2011. "Cognitive Task Demands, Self-Control Demands and the Mental Well-Being of Office Workers." *Ergonomics* 54 (9): 830–839.

Byers, J. C., A. C. Bittner, and S. G. Hill. 1989. "Traditional and Raw Task Load Index (TLX) Correlations: Are Paired Comparison Necessary?" In *Advances in Industrial Ergonomics and Safety*, edited by A. Mital, 481–485. London: Taylor & Francis.

Chen, S. M. 1996. "New Methods for Subjective Mental Workload Assessment and Fuzzy Risk Analysis." *Cybernetics and Systems* 27: 449–472.

François, M., D. Liévin, and M. et Mouzé-Amady. 2007. "Activité, charge de travail et stress du personnel navigant des compagnies aériennes: la situation dans les courts et moyens courriers." *Documents pour le Medecin du Travail* 111: 307–333. http://www.inrs.fr/accueil/dms/inrs/CataloguePapier/DMT/TI-TC-115/tc115.pdf

Gilles, M. A., G. Dietrich, K. Debuyser, J.-C. Guélin, and G. Didry. 2007a. "Variability of Motor Strategy as a Functional Index of Adaptation to a Working Task." In *Proceedings of the14th International Conference on Perception and Action (ICPA), Yokohama*, Japan, July 1–6, 181–182.

Gilles, M. A., G. Dietrich, J.-C. Guélin, and G. Didry. 2007b. "COM Mechanical Energy during Cyclical Working Task." In *Proceedings of European Workshop on Movement Science (EWOMS), Amsterdam*, The Netherlands, May 31–June 2, 128–129.

Gregoire, J., S. Tuck, R. L. Hughson, and Y. Yamamoto. 1996. "Heart Rate Variability at Rest and Exercise: Influence of Age, Gender, and Physical Training." *Canadian Journal of Applied Physiology* 21 (6): 455–470.

Hart, S. G. 2006. "NASA-Task Load Index (NASA-TLX); 20 Years Later." In *Human Factor and Ergonomics Society Annual Meeting Proceedings, General sessions*, October 2006, Vol. 50, No. 9, 904–908.

Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Worload*, edited by P. A. Hancock, and N. Meshkati, 139–183. Amsterdam: Elsevier.

Hoshikawa, Y., and Y. Yamamoto. 1997. "Effects of Stroop Color-Word Conflict Test on the Autonomic Nervous System Responses." *American Journal of Physiology, Heart and Circulatory Physiology* 272 (3): H1113–H1121.

ISO 10075-3. 2005. Ergonomic Principles Related to Mental Workload – Part 3: Principles and Requirements Concerning Methods for Measuring and Assessing Mental Workload.

Johnson, A., and A. Widyanti. 2011. "Cultural Influences on the Measurement of Subjective Mental Workload." *Ergonomics* 54 (6): 509–518.

Liévin, D., and M. et François. 1997. "Intégration d'une tâche de contrôle dans l'activité d'opérateurs de fabrication: le cas d'un atelier d'électronique." *Les Notes Scientifiques et Techniques de l'INRS* 153: 4–28.

Liu, T. S., and M. J. J. Wang. 1994. "Subjective Assessment of Mental Workload. A Fuzzy Linguistic Multi-Criteria Approach." *Fuzzy Sets and Systems* 62: 155–165.
Murofushi, T., and M. Sugeno. 1991. "A Theory of Fuzzy Measures: Representations, the Choquet Integral, and Null Sets." *Journal of mathematical Analysis and Applications* 159 (2): 532–549.
Noyes, J. M., and D. P. J. Bruneau. 2007. "A Self-Analysis of the NASA-TLX Workload Measure." *Ergonomics* 50 (4): 514–519.
Nygren, T. E. 1991. "Psychometric Properties of Subjective Workload Measurement Techniques: Implication for use in Assessment of Perceived Mental Workload." *Human Factors* 63 (1): 17–33.
Prade, H., A. Rico, M. Serrurier, and E. Raufaste. 2009. "Eliciting Sugeno Integrals: Methodology and a Case Study." In *Proceedings European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'09, Verona, Italy*, edited by C. Sossai and G. Chemello, July 1–3, 712–723. Springer, LNCS 5590.
Raufaste, E., and H. Prade. 2006. "Sugeno Integrals in Subjective Mental Workload Evaluation: Application to Flying Personnel Data." In *Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'06)*, 564–570.
Sokal, R. R., and F. J. Rohlf. 1995. "Biometry." In *The Principles and Practice of Statistics in Biological Research*. 3rd ed. New York: WH Freeman & Co.
Stroop, J. R. 1935. "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology* 18: 643–662.
Sugeno, M. 1974. "Theory of Fuzzy Integral and its Application.". PhD diss., Tokyo Institute of Technology.
Sugeno, M. 1977. "Fuzzy Measures and Fuzzy Integrals: A Survey." In *Fuzzy Automata and Decision Processes*, edited by M. M. Gupta, G. N. Saridis, and B. R. Gaines, 89–102. Amsterdam: Elsevier.
University of Kuopio. 2008. Kubios HRV 2.0. http://kubios.uku.fi/
Von Borell, E., J. Langbein, G. Després, S. Hansen, C. Leterrier, J. Marchant-Forde, R. Marchant-Forde, et al. 2007. "Heart Rate Variability as a Measure of Autonomic Regulation of Cardiac Activity for Assessing Stress and Welfare in Farm Animals – A Review." *Physiology and Behavior* 92 (3): 293–316.
Webber, C. L., and J. P. Zbilut. 1994. "Dynamical Assessment of Physiological Systems and States Using Recurrence Plot Strategies." *Journal of Applied Physiology* 76: 965–973.
Zbilut, J. P., N. Thomasson, and C. L. Webber. 2002. "Recurrence Quantification Analysis as a Tool for Nonlinear Exploration of Nonstationary Cardiac Signals." *Medical Engineering & Physics* 24: 53–60.

Appendix 1. Sugeno Integral modelling

In a finite evaluation setting where $N=\{1,\ldots,n\}$ indexes the set of criteria, a fuzzy measure is a set function from 2^N (power set of N, i.e. the set of all subsets in N) to [0,1] such that

(1) $\mu(\varnothing) = 0$, where $\varnothing$ is the empty set;
(2) $\mu(N) = 1$;
(3) if $A \subseteq B \subseteq N$, then $\mu(A) \leq \mu(B)$.

$\mu(A)$ and $\mu(B)$ can be viewed as the weights of importance of the sets of elements A and B, respectively.

A fuzzy integral associated with a membership function f from N over a non-fuzzy subset $A \subseteq N$ is defined as

$$S_\mu(f) = \max_{i=1,n} \min(f(\sigma(i)), \mu(F_{\sigma(i)})),$$

with $F_{\sigma(i)} = \{\sigma(i), \sigma(i+1), \ldots, \sigma(n)\}$, provided that N has been reordered through the permutation σ such that $f(\sigma(1)) \leq \ldots \leq f(\sigma(n))$. Here, the set N of criteria corresponds to the set of six ratings $(f_1, \ldots, f_6)$ used in NASA-TLX. In the classical NASA-TLX approach, the subjective workload $SW(f)$ is assumed to be a weighted sum of the f_i's where weights are elicited from the respondents. In this paper, we assume that for a set of respondents, there exists a Sugeno integral S_μ such that

$$\mathrm{SW}(f) \approx S_\mu(f),$$

where the fuzzy measure μ is unknown and has to be calculated. Leaving to the algorithm the task to determine μ, this means that we need only an external criterion (e.g. a global subjective rating GSW (f) *or* an objective, say physiological, measure) in addition to the six classical ratings, to bypass the PWT.[10]

Appendix 2. Practical guidelines for computing the Fuzzy-TLX

Since no public domain software currently exists, researchers have to develop their own version. We provide here the logic of the Fuzzy-TLX algorithm, which a professional programmer can follow to implement the method.

Data collection: With the six subjective ratings of the original NASA-TLX, at least one global measure must be provided by the participants to serve as an aggregation criterion. This may be a global subjective rating or some objective cue (a mean heart-rate parameter, a mean reaction time and so on). Nevertheless, this global criterion must be in the same range as the six ratings (e.g. 0–100). This can be straightforward when using a global subjective rating, but it will require a mathematical transformation for objective data measured in specific units.

Weight extraction: Input data are composed of observations, each having six classical ratings and a global criterion (c), all in the same range (e.g. 0–100). The extraction process seeks to define the intervals of possible values for the weights associated with each of the 64 (2^6) subsets of workload ratings {Ø, {1}, {2}, …, {6}, {12}, {13}, …, {56}, …, {123456}}. The weights range within the same bounds as the rating scales (here, from 0 to 100). At the beginning of the process, only two weights are known: $w(\text{Ø}) = 0$ and w

$(\{123456\}) = 100$.[11] Nothing is known about the 62 other weights, except their range (from 0 to 100). By definition, given an n-tuple of ratings, the Sugeno integral is the median of $2n - 1$ terms. Thus, in the case of Fuzzy-TLX, the terms of each observation are the $n = 6$ ratings and 5 weights, with c serving as an estimate of the 'real' Sugeno integral. Each new observation may provide information about some of the 62 weights to be identified. Given one observation, the six ratings are known. From the rating values in the observation, we can find which, among the global list of 62 weights, are the 5 weights relevant to this observation. Yet, we generally do not know the value of these 5 weights. Nevertheless, because the global criterion is a median, its location within the six ratings constrains the possible values of the 5 ongoing weights. When processing the whole data file, the algorithm capitalises the information brought by each observation to progressively restrain the intervals characterising the possible weights of the 62 subsets. For each observation, run the following steps:

Finding the five relevant subsets: First, sort the ratings $r(1)$ to $r(6)$. For example, let the six ratings be 22, 15, 56, 36, 41 and 55. Thus, $r(3) > r(6) > r(5) > r(4) > r(1) > r(2)$. Then we get $\{3\}$ as the first relevant subset, the second is $\{36\}$, the third is $\{356\}$ and so on until $\{13456\}$.[12] Taking into account the global criterion, c, which is the estimate of the Sugeno integral for this observation: here $c = 50$. Since the value of a Sugeno integral lies always in the boundaries of the individual ratings, we should have $\min(r(1), r(6)) \leq c \leq \max(r(1), r(6))$, here $15 \leq c \leq 56$. If c did not fit this compatibility constraint (i.e. if we had $c < 15$ or $c > 56$), the observation would be incompatible with any Sugeno integral family and therefore could not be used in the process.[13] Otherwise,

Locating c in the ranked list of terms: Because c is the median of 11 terms, 5 terms are above and 5 below c. Thus, by tallying the ratings above c, below c and equal to c, we can refine the intervals of possible values. Here, two ratings above c leave three slots for the weights. Four ratings below c leave only 1 weight.

Extracting information: Given the monotony property of fuzzy measures,[14] $w\{3\} \leq w\{36\} \leq w\{356\} \leq w\{3456\} \leq w\{13456\}$. Therefore, $w\{36\} = c = 50$. In the case of ties – between ratings, or between c and some ratings – we may learn less but the principle remains the same. There we can refine the range of possible values for those subsets. By monotony, we can also refine other embedded or embedding subsets. Since $w\{36\} = 50$, it follows that $w\{3\} \leq 50$ and $w\{6\} \leq 50$. On the other side, $50 \leq w\{136\}$, $50 \leq w\{236\}$, $50 \leq w\{356\}$ and so on.

Updating the fuzzy measure: If we know, from previous observations, that $10 \leq w\{16\} \leq 30$ and we learn thereafter that $w\{16\} \leq 25$, we can update our knowledge to $10 \leq w\{16\} \leq 25$. However, when the information brought by a particular observation becomes available, it must be checked whether this new information is compatible with all available intervals. Inconsistency arises when at least one interval does not overlap: e.g. we know that $10 \leq w\{16\} \leq 30$ and the new observation entails that $w\{16\} \geq 40$. In this case, the observation must be integrated to another consistent cluster of observations.

Note on the introduction of observations: The method described above is sensitive to order effects. To minimise biases, we used a minimum specificity principle: i.e. examining the amount of change potentially introduced by each observation and always choosing the observation that produces the lowest change. Another introduction strategy, which uses simulated annealing techniques to build data clusters consistent with the same Sugeno integral family, can be applied (Prade et al. 2009).

Final Fuzzy-TLX: Once all observations have been reviewed, the weights of their relevant cluster can be used to compute the two Fuzzy-TLX values: SugMin (from all the min values of possible weights) and SugMax (from all the max values of possible weights).